Preprint
UCRL-JC-149741

Biosignatures of Pathogen and Host

J. P. Fitch, et al

This article was submitted to
Workshop on Genomic Signal Processing and Statistics,
Raleigh, North Carolina
10/12/02 – 10/13/02

August 27, 2002

U.S. Department of Energy
Lawrence
Livermore
National
Laboratory

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

BIOSIGNATURES OF PATHOGEN AND HOST

J.P. Fitch, B.A. Chromy, C.E. Forde, E. Garcia, S.N. Gardner, P.P. Gu, T.A. Kuczmarksi, C.F. Melius, S.L. McCutchen-Maloney, F.P. Milanovich, V. L. Motin, L.L. Ott, A.A. Quong, J.N. Quong, J.M. Rocco, T. R. Slezak, B.A. Sokhansanj, E. A. Vitalis, A. T. Zemla, and P. M. McCready

Chemical and Biological National Security Program
Lawrence Livermore National Laboratory

ABSTRACT

In information theory, a signature is characterized by the information content as well as noise statistics of the communication channel. Biosignatures have analogous properties. A biosignature can be associated with a particular attribute of a pathogen or a host. However, the signature may be lost in backgrounds of similar or even identical signals from other sources. In this paper, we highlight statistical and signal processing challenges associated with identifying good biosignatures for pathogens in host and other environments. In some cases it may be possible to identify useful signatures of pathogens through indirect but amplified signals from the host. Discovery of these signatures requires new approaches to modeling and data interpretation. For environmental biosignal collections, it is possible to use signal processing techniques from other applications (e.g., synthetic aperture radar) to track the natural progression of microbes over large areas. We also present a computer-assisted approach to identify unique nucleic-acid based microbial signatures. Finally, an understanding of host-pathogen interactions will result in better detectors as well as opportunities in vaccines and therapeutics.

1. INTRODUCTION

The motivation for our research is to establish diagnostics for infectious diseases that are highly specific, easy to use, rapid, and useful in a variety of media and applications. We have characterized the progression of signatures as

I. Empirically discovered signatures,
II. Signatures located on the pathogen genome—i.e., which gene(s) contains the signature, and
III. Association of the signature with a function of the pathogen and/or the host response.

Historically, an Edisonian approach to diagnostic signature discovery has been utilized that leveraged intuition from microbiological experts, immune responses (e.g., antibodies), and limited tools for comparisons among species. In attempting to duplicate and extend the traditional approach to target other infectious agents we frequently found a very low return on investment. It was not uncommon to test over 1,000 "promising" diagnostics to find that none of them could identify all strains of the target pathogen or they would produce false positives in the presence of similar (but different and often non-pathogenic) microbes. The first challenge we faced was to invent a reliable source of biosignatures for many pathogens.

Assuming a useful biosignature for the target pathogen exists (Type I signature), we also have a need to push the development towards Types II and III signatures in order to anticipate when the diagnostics may fail and to provide targets for inhibitors, vaccines and treatments. A useful test case is influenza. Each year, this virus returns to the human population with a modulated genome. Association of the biosignature with mechanism/function provides benefit above and beyond detection. It appears that a Eukaryotic host responds to pathogens in ways that may amplify the number and volume of potential biosignatures [1]. Finally, it is essential that the pathogen can be found (if present) in complex environments including animals, plants, air, soil, and water. This is a concentration or, in the traditional information theory lingo, signal to noise ratio issue. The number of background microbes as well as chemical and physical inhibitors can interfere with successful application of the diagnostic sensor—even one based on an excellent biosignature. Because of this, it is important to understand where to look in the environment for signature-similar microbes and use these to optimize pathogen signatures.

In this paper, we summarize our approach to computer-assisted design of nucleic acid based diagnostics with additional details available in Fitch [2]. Preliminary results on *Yersinia pestis* as a model organism for studying mechanisms of pathogenicty, host response, and, therefore, Type III signatures are presented. We include a discussion of a subset of the many potential approaches to modeling, simulating, and analyzing genomic and proteomic data for the purpose of improved Type III signatures. We refer to this scientific endeavor as *Path-omics:* the comprehensive (omics = total mass of knowledge) study of disease and causative mechanisms (path = disease and/or disease causing). Finally, we

provide a context for environmental monitoring for pathogens and proposed approaches to combining sensor measurements with computer models. All of these closely related applications are described in the context of a systems approach to biological applications.

2. SYSTEM-LEVEL HYPOTHESIS TESTING

Advances in experimental techniques are generating vast amounts of data including nucleic acid sequence, genes, transcript profiles of mRNA, protein expression levels, post-translational modifications, and *in vivo* experiments [3, 4]. The data provide us with the opportunity to apply an integrated systems approach for exploring biological pathways (e.g., signaling, transcription, translation, secretion, adhesion) and the emergent behavior of the system. This systems approach allows the study of the interactions of a complete set of objects, e.g. genes or proteins, rather than simple binary interactions between two objects. To systematically elucidate all of the interactions between the objects would require a combinatorially large number of experiments. Even with the advent of high-throughput experimental techniques, it is not possible to perform all experiments to explicitly identify each and every reaction. Fortunately, it is recognized that not all interactions are active in the system [5] and performing the correct subset of experiments will provide the needed information. The challenge is how to define the appropriate experiments and then analyze the resulting large amounts of data produced in the high-throughput model? The answer to this question is an integrated systems approach coupling the experimental regime with theory and modeling. The computational model can eliminate unnecessary experiments that would provide duplicative data. Furthermore, by integrating the modeling with the experiments, high throughput analysis of key biochemical species can be maintained in less time than manual analysis.

Hypothesis testing is a basic tenet of science. But with a large number of objects (e.g. over 4,500 genes in *Yersinia pestis*), conceptualizing the outcome of the hypothesis requires a computational model. In the systems biology paradigm, hypothesis testing involves global analysis—this requires the development of novel computational algorithms. We are therefore taking an integrated experimental and computational approach to hypothesis testing at the systems level.

This integrated approach is illustrated in Fig. 1. Experimental data is first used to generate a model of the system. This, in itself, is a difficult inverse problem, requiring iteration. In the next step, the model, which is consistent with existing experimental data, is used to predict new behavior of the system. In order to test the hypothesis, i.e. the predicted behavior, from the model, new experiments are designed to test the hypothesis. The new experiments themselves may be quite diverse, corresponding to gene expression, protein or metabolite concentrations, or overall cellular behavior. Finally, the newly designed experiment generates new experimental data, which must be integrated into the existing model to generate the improved model. This leads to an iterative, hypothesis driven approach to data-driven model development.

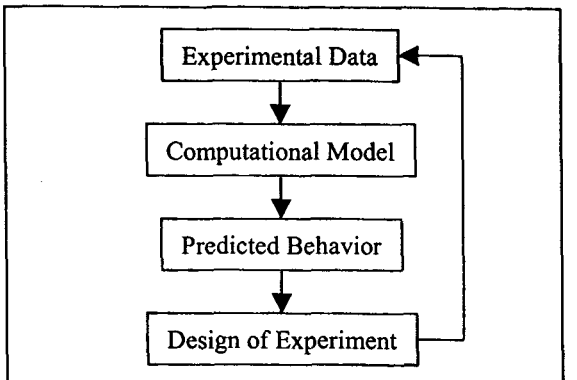

Fig. 1 The role of computer models in systems-level hypothesis testing biology. Determining an appropriate model as well as the correct parameters for the model is a difficult inverse problem.

3. NUCLEIC ACID PATHOGEN SIGNATURES

In August 2000, a variety of traditional wet-lab development methods were failing to yield robust pathogen signatures. In testing over 1,000 candidates, all cross-reacted with neighbor species commonly found in background environments within the continental US. This prompted a quick and dirty test, using the BLAST [6] algorithm to compare the entire genome of the target pathogen against all the microbial genomes then in Genbank (about 85). Several giga-bytes of raw BLAST output were parsed and all exact match regions of the pathogen genome with another genome were removed or "masked out" of the pathogen genome. All remaining pathogen sequence was then considered "potentially unique" to the target pathogen.

The MIT Primer3 program [7] was used to design potential signature primer pairs on each fragment of sufficient length. Nearly 4,000 signature candidates on the ~5Mbase pathogen genome resulted from this crude BLAST parsing effort. We chose 400 at random, ordered the PCR primer oligos, and screened against the panel of near-neighbors that had eliminated all 1,000 of the wet-lab candidates. Several dozen of the candidate signatures passed this screening, and 4 survived rigorous testing of the complete set of DNAs used to ensure a high degree of specificity and have been successfully fielded. We realized that a threshold had been crossed: it was now

feasible to apply whole-genome analysis techniques to the field of DNA signature development. This was our first "systems" approach. We were encouraged by the initial success of the crude BLAST-parsing demonstration. However, we realized that a lot needed to be done to make even a modest prototype of an automated DNA signature generation pipeline.

Length of DNA sequence affects the scalability of DNA signature development algorithms. Techniques that may be computationally feasible for small viral genomes often fail on the genomes of bacteria that may be 1,000 times larger. These economic and scientific realities of DNA signature construction influenced our signature development process. The pipeline has two entry points, nominally one for viral and one for bacterial pathogens. The viral pathogen entry point can also be used as a last resort for bacterial genomes where only a few gene fragment sequences are available.

Virus strain genomes are pre-processed to extract a single "consensus gestalt" genome. Bacterial genomes are assumed to have (at best) one completed genome. In either case, we employ an efficient algorithm to compare the pathogen target against all other sequenced bacterial and viral genomes, "masking out" in the pathogen all DNA sequence over K bases in length that "match" (exactly or close enough for PCR) one or more of the other non-target genomes. We then mine the remaining "candidate-unique" DNA to search for suitable signature primers. These are checked electronically, stored in a database, and a subset is ordered using a computer-generated procurement process. Upon delivery from an external vendor, they are screened in the wet laboratory against the target DNA and against a panel of near-neighbors and environmental backgrounds. Successful ones are sent to the CDC (or other collaborator) for additional testing and validation. Validated assays are utilized in both public health and bio-defense applications.

See Nicholas [8] for a brief comparison of MSA algorithms. Unfortunately, many of the MSA algorithms and codes do not scale well for even modest genome lengths or for enough strains in the alignment. We located a very good MSA algorithm called DiAlign [9] that scales better than any other we tested.

Assuming that MSA can be completed successfully, we then process the aligned genomes into a "consensus gestalt" sequence. Most simply stated, this places a "dot" (period) in all positions of the output alignment if the bases in that position are not in agreement. The base is capitalized if all the bases in that position match and the position is part of at least an 18 base consecutive run of matches. Otherwise, the base is lower-case (match at that position but no run of 18 matches). An 18 position consecutive run of bases is used because this is the minimum length for a forward or reverse PCR primer.

Bacterial pathogens take a different route through the candidate signature pipeline. Even in the (rare today) case where more than one full-length bacterial genome is available, current MSA algorithms cannot handle input of that length. In addition, bacterial genomes violate the subtle assumptions of colinearity that underlie most MSA algorithms. Colinearity is violated due to real biological issues including genome rearrangements, gene duplication, and other events.

The most efficient algorithms for finding sub-strings that match between two inputs are known as "suffix-tree" algorithms [10]. We have utilized an extremely efficient suffix-tree implementation known as *mkvtree/vmatch* [11]. Our collaborator, S. Kurtz, added several features that met application-specific needs for signature development, and supplied the project with numerous binary executable updates and performance enhancements.

As our "pipeline" for nucleic acid signature discovery matures we are looking to use improved multi-sequence alignment tools as well as extend the approach to Type II and III signatures where the function of the sequence is utilized or even to design binding affinity assays.

4. BACKGROUNDS

Biological backgrounds studies are motivated to

1) Establish natural pathogen concentrations in the environment for environmental monitoring (e.g., deployments) and for future reference in forensics/attribution and restoration activities (e.g., cleaning up to the pre-existing environmental standard).

2) Utilize endemic microbes to track mobility of microbes in environments that are perceived to be potential terrorist targets. This is essentially an emperical threat assessment that can be used to pre-qualify modeling and simulation as well as place and elements of a sensor network.

To achieve these two goals requires a variety of improvements in sample collection and processing including enhanced throughput. The use of broad spectrum detectors (e.g., the 16S ribosomal DNA chip [12]) and microbe-specific signatures enable the second goal. Endemic microbes may come from a variety of sources that are representative of point and line releases including manmade, agricultural, and geographical—e.g., ponds.

We are developing a collection network that is densely sampled in both time and space. Aerosol and soil samples will be collected and processed using 16S analysis as well as specific PCR primers. Meteorological data is also needed to incorporate dispersion into the models. Several approaches to data analysis are possible: a Bayesian approach [13], a matched filter / optimal linear estimator similar to tomography and radar imaging [14],

and iteration between data and forward models using atmospheric propagation algorithms like those at NARAC.

This task will help validate atmospheric models and establish a method for empirical measurements in an urban environment with less environmental impact than a controlled release experiment. Preliminary evidence that endemic microbes can be tracked and modulated in the environment exists. It is not yet apparent from this data whether a decrease, increase or displacement was the cause of the modulation of the environmental microbes.

5. THE NEXT LEVEL SYSTEM

The computational challenge is the inverse problem of developing the model from experimental data. As an example, determination of the genetic network architecture from microarray data has two requirements. The first requirement is to find a mathematical description of the interactions between genes. It would seem natural to treat the genes as biomolecules and the expression levels as chemical concentrations. The coupling between genes and their time-evolution could then be treated as a set of coupled ordinary differential equations. However, the experimental data obtained does not lend itself to this treatment because the level of precision is insufficient to determine the parameters of a continuous numerical model. The fuzzy logic approach is one mathematical description used to represent this type of experimental data [15, 16]. Fuzzy logic provides the framework for the mathematical manipulation of imprecise quantitative data such as that generated from microarray data.

The second requirement is to develop algorithms that determine the parameters of the mathematical model that describes the interactions and evolution of the system. In this particular example of gene network architecture, this requires algorithms that determine how genes within a set interact with each other and how these sets interact. Just as importantly, these algorithms will determine which genes do not interact with each other. This is a technically difficult problem because of the combinatorially large number of interactions in an entire genome. An exhaustive search is not possible so one must resort to more efficient methods. Possibilities include genetic algorithms and programs, linear optimization techniques, neural networks, and stochastic methods. Finally, the models must be extensible to larger volumes and types of biological data.

6. ACKNOWLEDGEMENTS

The authors would like to thank our many collaborators in public health. Special thanks to the BASIS team that has translated the science into deployable reality. BASIS is led by W. Davidson at Los Alamos and by D. Imbro at Lawrence Livermore National Laboratories. This work was sponsored in part by LLNL LDRD SI investments under the Pathogen Pathway and *Pathomics* Projects and the DOE/NNSA Chemical & Biological National Security Program. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

7. REFERENCES

[1] Q. Huang, *et al*, "The Plasticity of Dendritic Cell Responses to Pathogens and Their Components," *Science*, vol. 294, pp. 870-875, Oct. 26, 2001.

[2] J.P. Fitch, *et al*, "Rapid Development of Nucleic Acid Diagnostics," *Proc. IEEE*, Nov. 2002.

[3] S. Tavazoie and G.M. Church, "Quantitative Whole Genome Analysis Of DNA-Protein Interactions By In Vivo Methylase Protection In *E. coli*," *Nature Biotechnology*, vol. 16, pp. 566-71, 1998.

[4] R.J. Cho, *et al.*, "A Genome Wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.

[5] P.M. Gleiss, P.F. Stadler, A. Wagner, and D.A. Fell, "Relevant Cycles in Chemical Reaction Networks," *Adv. Complex Systems*, vol. 4, pp. 207-226, 2001.

[6] National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST®) http://www.ncbi.nlm.nih.gov/BLAST/

[7] S. Rozen, H.J. Skaletsky, 1996 & 1997, www-genome .wi.mit.edu/genome_software/other/primer3.html

[8] H. B. Nicholas, Jr., *et al*, "Strategies for multiple sequence alignment," *BioTechniques*, vol. 32, No. 3, pp. 572-591, March 2002.

[9] "DiAlign 2: A novel algorithm for pairwise as well as multiple alignment of nucleic acid and protein sequences" http://www.gsf.de/biodv/dialign.html

[10] P. Weiner, "Linear pattern matching algorithms," *Proc. 14th IEEE Annual Symp. Switching and Automata Theory*, pp. 1-11, 1973.

[11] S. Kurtz, *mkvtree/vmatch*, obtained by request http://bibiserv.techfak.uni-bielefeld.de/vmatch/

[12] K.H. Wilson, et al, "High density microarray of small subunit ribosomal DNA probes," *Appl. Env. Micro.*, vol. 68, No. 5, May 2002.

[13] Van Trees, H.L., *Detection, Estimation, and Modulation Theory, Part I*, John Wiley & Sons, 1968.

[14] Fitch, J.P., *Synthetic Aperture Radar*, Springer-Verlag, 1988.

[15] P.J. Woolf, Y. Wang, "A Fuzzy Logic Approach to Analyzing Gene Expression Data," *Physiol.Genomics*, vol. 3, 9-15, 2000.

[16] B.A. Sokhansanj, J.P. Fitch, "Interpreting Microarray Data to Build Models of Microbial Genetic Regulation Networks," *SPIE Photonics West 2002 (BIOS)*, San Jose, CA, Jan. 19-25, 2002.